Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования «Сибирский федеральный университет»

Институт фундаментальной биологии и биотехнологии

ДИПЛОМНАЯ РАБОТА

Чернышова Анна Игоревна

СВЯЗЬ СТРУКТУРЫ И ТАКСОНОМИИ НА ПРИМЕРЕ ХЛОРОПЛАСТОВ

Студентка ФБ12-01Б 041202804 _____ А. И. Чернышова

Научный руководитель _____ М. Г. Садовский

Красноярск 2016

Оглавление

Введение

Изучение биологических макромолекул является одной из центральных проблем современной биологии, биотехнологии, биоинформатики и других смежных наук. Большое внимание привлекает к себе выявление и описание структурированности во множестве генетических данных [1–3,5,6]. Настоящая работа посвящена выявлению связи между структурой и таксономией на примере геномов хлоропластов. Данные генетические объекты были выбраны не случайно: прежде всего они обладают небольшим размером. Хлоропласты — внутриклеточные органоиды (пластиды) растений, в которых осуществляется фотосинтез. Их размер колеблется от 4 до 10 мкм, а число составляет от 20 до 100 на клетку [4]. Во-вторых, с точки зрения частотных словарей хлоропласты более менее однородны внутри, что так же является одной из причин выбора данных структур.

Цель данной работы — выявление, описание и анализ связи между структурой и таксономией геномов хлоропластов. Для достижения цели необходимо решить следующие задачи:

- определить, что такое структура и таксономия в рамках данного исследования;
- определить, насколько разные геномы оказываются близкими по структуре и формируют ли они кластеры;

- выделить такие кластеры;
- изучить структуру кластеров в терминах случайности и неслучайности их состава и связи между такими кластерами.

Основные результаты работы были представлены на международных и Всероссийских конференциях:

- Международная научная конференция «Проспект Свободный — 2016», 2016, г. Красноярск, устный доклад «Построение связи между структурой и таксономией геномов хлоропластов сосен»;
- МНСК 2015, г. Новосибирск, устный доклад «Проявление синхронизации в эволюции геномов растений»;
- Всероссийский семинар по нейроинформатике, 2014 г., г. Красноярск, устный доклад «Проявление синхронизации в эволюции геномов растений и их хлоропластов»;
- IWBBIO 2015, Granada, Spain, устный доклад «Genome Structure of organelles strongly relates to taxonomy of bearers»;
- XIV Межд. ФАМ-конференция, Красноярск, 2015, устный доклад «Синхронизация эволюции растений и их хлоропластов»;
- ECCS 2014, Lucca, Italy, устный доклад «Revealing the Relation Between Structure of Chloroplast Genomes and Host Taxonomy»;
- BioMath 2014, Będlevo, Poland, устный доклад «Very high synchrony in evolution of organelles and host genomes»;
- МНСК 2014, г. Новосибирск, устный доклад «Выявление связи между структурой и таксономией геномов хлоропластов».

Основные результаты работы также опубликованы в 10 публикациях (см. список литературы).

Глава 1. Основные понятия и определения, обзор литературы

1.1 Постановка задачи

В данной работе представлены результаты по изучению связи между структурой и филогенией геномов хлоропластов. Структуру определяли по частотному словарю, построенному по геному хлоропласта. В свою очередь, таксономию определяли по соматическому геному. Кластеризацию проводили методом динамических ядер, в программе VidaExpert. Основная задача работы — выявление связи между классами, получившимися методом динамических ядер, и их видовым (такосномическим) составом. Как было установлено, кластеры действительно образовывались неслучайно: в одни и те же классы, как правило, попадали филогенетически близкие виды.

Предметом нашего анализа являются статистические свойства молекул ДНК. Абстрагируясь от химических свойств этих молекул, будем рассматривать такие молекулы как символьные последовательности из 4-х буквенного алфавита $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$; число символов N в последовательности будем считать её длиной. Будем предполагать, что никаких других символов в последовательности не содержится (см. п. 2.1.1). Кроме того, все последовательности будут считаться связными (т. е. не содержащими пробелов).

1.1.1 Частотные словари и их свойства

Рассмотрим символьную последовательность из четырёхбуквенного алфавита $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, соответствующую той или иной генетической последовательности; всюду далее мы будем рассматривать только последовательности геномов хлоропластов. Любую связанную символьную последовательность длины q будем называть словом (соответствующей длины). Частотный словарь W_q представляет собой список всех слов $\omega = \nu_1\nu_2\ldots\nu_{q-1}\nu_q$ подряд идущих q символов с указанием их частот.

Частота f_ω данного слова — это отношение числа копий n_ω данного слова $\omega = \nu_1\nu_2\ldots\nu_{q-1}\nu_q$ к общему числу всех слов. Всюду далее в настоящей работе будут рассматриваться исключительно частотные словари W_3 (толщины 3) символьной последовательности, соответствующей ДНК — словари триплетов. Следует отметить, что в данной работе разница между триплетом и кодоном содержательна. Дело в том, что триплет представляет собой любую тройку подряд идущих символов в последовательности. Кодон является триплетом, но подсчёт кодонов ведётся с условием того, что

а) два кодона не пересекаются, но примыкают друг к другу без разрыва;

б) обычно под кодоном понимается триплет, который занимает вполне определённое место в последовательности — отсчёт кодонов ведётся от стартового;

в) каждый кодон (в отличие от триплета) кодирует какую-либо аминокислоту.

1.1.2 О расстоянии между словарями

Как только символьные последовательности (геномы хлоропластов в нашем случае) оказываются преобразованными в частотные словари, появляется возможность определить близость двух геномов естественным образом — как близость двух точек в метрическом пространстве, определяемая в той или иной метрике. В настоящей работе всюду использовалось Евклидово расстояние между двумя частотными словарями триплетов:

$$\rho\left(W_3^{(1)}, W_3^{(2)}\right) = \sqrt{\sum_{i=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_i^{(1)} - f_i^{(2)}\right)^2}. \tag{1.1}$$

Индекс i пробегает все триплеты от AAA до TTT. Евклидово расстояние — наиболее распространенное расстояние; возможен выбор и других метрик, однако их использование, а также сравнение результатов кластеризации (см. раздел 2.1.4, стр. 16), которые могли быть получены для разных метрик, не входило в задачи настоящей работы.

1.1.3 Об исключении триплета

Всякий частотный словарь отображает геном в 64-мерное метрическое пространство; однако сумма частот всех триплетов

$$\sum_{i=\mathsf{AAA}}^{\mathsf{TTT}} f_i = 1 \tag{1.2}$$

накладывает сильную линейную связь: на самом деле каждая точка лежит на линейном подпространстве коразмерности 1 в линейном пространстве размерности 64. Тем самым, для исключения влияния этой линейной связи один триплет в целях дальнейшего анализа необходимо исключить.

Формально можно исключать любой триплет, в нашей работе мы исключали не любой, а вполне конкретный. Обсуждение правила выбора такого исключаемого триплета см. в разделе 2.1.2 (стр. 14).

1.1.4 Слоистые графы

Дадим определение графу: пусть имеется множество точек V, называемое вершинами графа. Отрезки, соединяющие вершины, назовём рёбрами графа, а совокупность множества вершин и рёбер — графом

$$G = G(V) \tag{1.3}$$

с множеством вершин V. Каждая конкретная пара называется ребром графа, вершины a и b называются концевыми точками или концами ребра E. Если порядок расположения двух концов ребра несущественен, т. е. если

$$E = (a, b) = (b, a)\,, \tag{1.4}$$

то E есть неориентированное ребро, если же этот порядок существенен, то E называется ориентированным ребром. Соответственно, если у графа каждое ребро не ориентировано, то он неориентированный, если ориентированы все его рёбра, то ориентирован, если же у графа есть ориентированные и неориентированные рёбра, то такой граф является смешанным. Также на множестве графов выделяют нуль-графы и полные графы. В первом случае граф состоит только из изолированных вершин. Во втором рёбрами являются всевозможные пары для двух различных вершин a и b из V (случай неориентированного полного графа). В ориентированном полном графе имеется пара

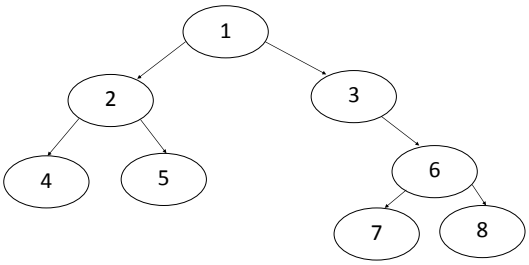

Рис. 1.1. Пример графа, являющегося деревом.

рёбер, по одному в каждом направлении, соединяющие любые две различные вершины a и b. Если любая пара вершин связана хотя бы одним маршрутом (набором рёбер, начинающимся в заданной вершине и заканчивающейся в другой заданной вершине), то такой граф называется связным.

Для целей настоящего исследования нам понадобится специальный случай графа: слоистые графы. Слоистым графом будем называть граф $G(V)$ такой, что все его вершины разбиты на непересекающиеся подмножества вершин, называемые слоями. Точнее, слоем будем называть подмножество $G^{(l)}(V)$ вершин графа $G(V)$, $1 \leqslant l \leqslant l^*$ таких, что эти подмножества считаются упорядоченными:

$$G^{(1)}(V) \prec G^{(2)}(V) \prec \cdots \prec G^{(l-1)}(V) \prec G^{(l)}(V)\,, \tag{1.5}$$

где символ $\prec$ обозначает отношение упорядочения. Следует подчеркнуть, что данное отношение не является естественным, но накладывается на множества вершин графа $G(V)$ по соглашению.

Слоистый граф — это такой граф, в котором рёбрами соединены вершины, принадлежащие соседним слоям, и только они. В нашем случае мощность слоя (число вершин в слое) монотонно возрастает с ростом порядка слоя. В зависимости от структуры слоистого графа он может предщставлять собой два предельных случая: в первом случае слоистый граф сводится к дереву, во втором — он является послойно полносвязным. Этот второй случай означает, что каждая вершина из слоя $l-1$ соединена с каждой вершиной из слоя l [7].

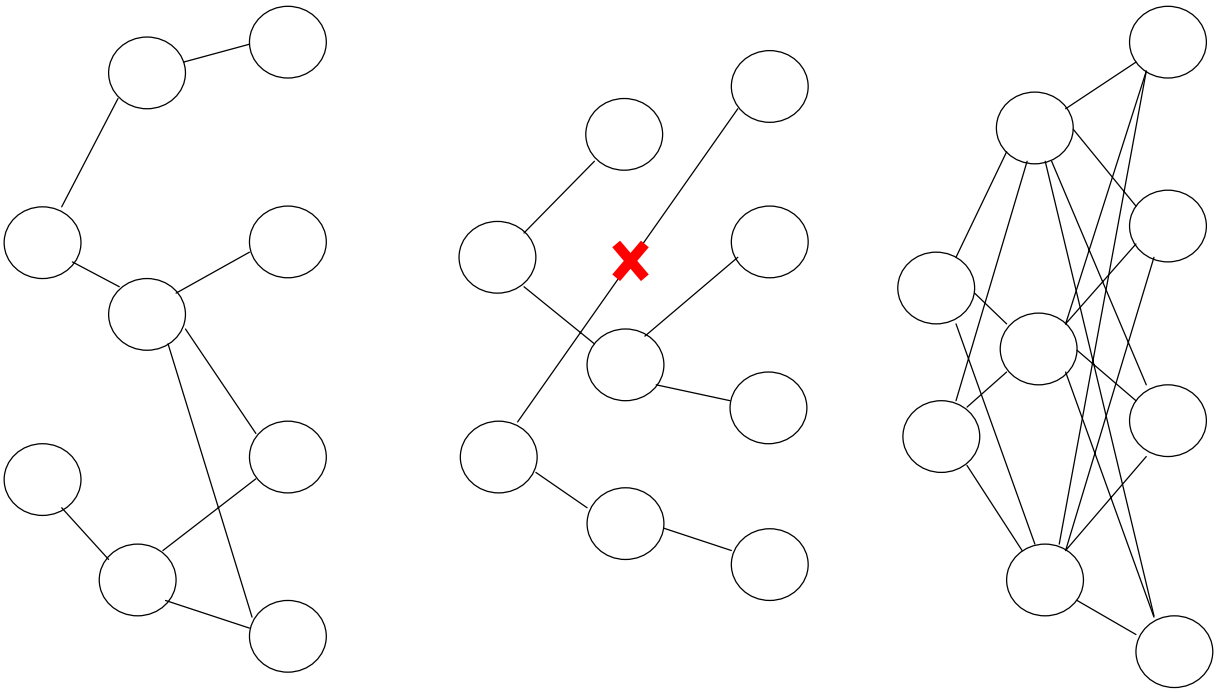

Рис. 1.2. Примеры графов. Графы справа и слева являются слоистыми, в центре слоистым не является.

Пример полносвязного графа показан на рис. 1.2 (справа), неполносвязного — на рис. 1.2 (слева).

1.2 Обзор литературы

В современном мире изучению структуры и свойств биологических мактромолекул посвящено бесконечное множество работ. Наиболее актуальными темами исследования являются описание таксономии растений, связь структуры и функции. К примеру, в работе [8] проведён сравнительный анализ данных о вирусах и разработана система таксономии виросов в классическом виде. В свою очередь, в работе красноярского учёного Бадмаева Б.Б. [9] проведено сравнение таксономической структуры кормовых растений заподного забайкалья. Можно долго перечислять подобные работы, ибо по данной

тематике их огромное количество. Для настоящей работы больший интересн вызвали работы [15, 16], а также [18] и более поздние версии [10, 11]. Первые две ([15, 16]) посвящены исследованиям классификации нуклеотидных последовательностей бактериальных РНК. По итогу работ было показано существование корреляции между таксономическим положением носителей и информационной структурой нуклеотидных последовательностей. В работе [18] предложен новый метод описания и выделения в целостных нуклеотидных последовательностях слов, которые обладают высокой информационной ценностью. В свою очередь, в статье [11] показано полное описание современной такосномии растений, что весьма пригодилось в настоящей работе. Кроме этого, также было интересно изучить работу [14], в которой представлены основные результаты по выявлению связи структурой и филогенией митохондриальных геномах. Результаты этой работы можно было сравнить с результатами настоящей работы. Оказалось, что как в митохондриях, так и в хлоропластах геномы розбиваются на кластеры не случайным образом.

Глава 2. Материалы и методы

2.1 База данных

Весь генетический материал для исследования был взят на сайте *www.ebi.ac.uk/genomes/organelles* в базе EMBL-банка (релиз от 2 мая 2015 года) и в базе GenBank (релиз от 24 июня 2015 года). В данном релизе представлено 937 геномов хлоропластов. База содержала геномы из разных сайтов, в результате чего, большую часть данных составляли геномы либо одинаковых, либо очень похожих видов растений. Из-за этого кластеризация методом динамических ядер может показать не верный результат: в кластеры будут объединяться одинаковые виды и связь выявлена не будет между структурой и таксономией. В связи с этим, исходную базу данных индексировали. Кластеризация проводилась на 718 геномах.

2.1.1 Соглашение о лишних символах

Как уже было сказано выше, непрерывная символьная последовательность из 4-х буквенного алфавита $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ длиной N соответствует генетической молекуле (или иначе генетическому тексту). Реальные последовательности, депонированные в EMBL–банке, могут содержать символы, не входящие в алфавит $\aleph$. В данной работе такие символы игнорировались, а последовательность конкатенировалась: фрагменты, содержащие лишние

символы, удалялись, и сама последовательность «состыковывалась». В результате последовательность длины N состоит лишь из символов $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

2.1.2 О выборе триплета

Так как сумма частот всех триплетов в словаре равна 1, то один из 64-х триплетов необходимо исключить. Выбор триплета, который необходимо исключить, зависит от варианта распределения. Бывают случаи, когда частоты отличаются на порядки. Здесь следует исключать триплет с максимальной частотой. Другой подход заключается в исключении триплета, вносящего наименьший вклад в различение геномов.

Действительно, пусть для некоторого множества геномов стандартное отклонение частот триплета ω^* равно нулю; это означает, что частота этого триплета во всех рассматриваемых геномах одинакова. Что, в свою очередь, означает, что данный триплет не вносит никакого вклада в различимость геномов. Данное наблюдение позволяет сформулировать правило исключения триплета: следует исключать тот триплет ω^*, для которого величина стандартного отклонения частот, определяемых в целом по базе, минимальна. В настоящей работе большого разброса частот не было и исключался триплет с минимальным стандартным отклонением. Как видно из таб. 2.1,

Таблица 2.1. Наименьшие и наибольшие стандартные отклонения; σ — стандартное отклонение.

$\min\{\sigma\}$	
GCG	0,004549
CGC	0,004629
GGC	0,006401
GCC	0,006447
CGG	0,006859

$\max\{\sigma\}$	
TTT	0,0437977
AAA	0,0427006
AAT	0,0345516
ATT	0,0345314
ATA	0,0288119

исключать следует триплет GCG, для случая того набора геномов, который рассматривается в работе.

2.1.3 О программе ViDaExpert

Исследование кластеризации геномов хлоропластов в пространстве частот триплетов проводились методом динамических ядер с использованием специальной программы *VidaExpert* [12]. Эта же программа использовалась для визуализации полученных результатов.

VidaExpert — одна из свободно распространяемых программ, которая позволяет представить наглядно многомерные данные и результаты их обработки. Программа VidaExpert имеет внутреннюю иерархию объектов. С какими-то работает исследователь, а какие-то являются контейнерами, которые упорядочивают объекты исследования. К примеру, контейнером верхнего уровня является проект. Он содержит в себе совокупность набора данных и карт, называемых сценарием. Каждый сценарий в свою очередь содержит таблицу и набор объектов данные. Таблица хранит исходную информацию, данные содержат числовой массив всех значений выбранных признаков. Объект типа данные создаётся с помощью объекта таблица. На основе данных пользователь создаёт карту, которая содержит всю исходную информацию о положении узлов сетки в пространстве и о способе её доопределения до многообразия. Чтобы визуализировать данные на основе карты создаются слои, которые содержат всю необходимую информацию для отрисовки на экране информационного слоя.

В VidaExpert 4 вида слоёв: слой точек данных, слой сетки, слой раскрасок, слой объектов. Каждый слой характеризуется своим видом:

– вид на координатной плоскости;

– вид на плоскости первых двух главных компонент;

– вид во внутренних координатах карты — простая развёртка карты;

– вид во внутренних координатах карты — нелинейная развёртка карты.

Здесь простая развёртка карты — равномерная сетка узлов, у которой точки данных размещены в соответствии с их проекциями на карту, а нелинейная развертка представляет собой криволинейную сетку.

2.1.4 Метод динамических ядер

Алгоритм кластеризации методом динамических ядер применяется к объектам, которые представляются точками в d-мерном векторном пространстве (в нашем случае 63-мерное метрическое пространство). Метод динамических ядер разделяет множество F на K кластеров точек, при этом каждая точка попадает только в один определенный класс. Как правило, в подобных алгоритмах кластеризации точки группируются по показателю близости их друг к другу. В нашем случае для реализации метода динамических ядер такой мерой близости было Евклидово расстояние. Изложим работу метода подробнее. На первом шаге некоторое множество F, состоящее из объектов, делят произвольно на K классов и каждую точку приписывают к соответствующему классу. Затем выполняется следующий алгоритм:

– для каждого класса определяют динамическое ядро — среднее арфиметическое значение частот каждого из триплетов:

$$c^{(j)}_{\nu_1\nu_2\nu_3} = \frac{1}{M^{(j)}} \sum_{i=1}^{M^{(j)}} \left(f_{\nu_1\nu_2\nu_3}(i) \right). \tag{2.1}$$

Здесь индекс i $(1 \leqslant i \leqslant M^j)$ перечисляет элементы класса. Среднее

арифметическое определяется для каждого триплета;

– затем для каждой точки и для каждого класса вычисляется расстояние между этой точкой и каждым из центров класса;

$$\rho^{(i)} = \sqrt{\sum_{\nu_1\nu_2\nu_3} \left(f_{\nu_1\nu_2\nu_3}(l) - c^{(i)}_{\nu_1\nu_2\nu_3} \right)^2}. \qquad (2.2)$$

Индекс i теперь перечисляет все полученные классы ($1 \leqslant i \leqslant K$), а индекс l перечисляет все точки множества, вне зависимости от того, к какому классу она принадлежит;

– если точка, первоначально приналежавшая классу K_1, оказывается ближе к центру класса K_2, то принадлежность этой точки к классу меняется (в рамках нашего примера с класса K_1 на класс K_2);

– положения центров вновь полученных классов пересчитываются и вся процедура продолжается до тех пор, пока переход точек из класса в класс не прекратится [19, 21].

Построение кластеризации методом динамических ядер зависит от нескольких факторов. Прежде всего, заранее неизвестно на какое максимальное количество классов следует разбивать анализируемое множество. В рамках настоящей работы не проводилось никаких специальных исследований, направленных на выявление оптимального числа классов, получаемых методом динамических ядер. Количество классов бралось равным восьми.

Устойчивость построения классификации также была не очевидна. Поскольку построение кластеризации методом динамических ядер всегда начинается со случайного разбиения объектов на классы, постольку может случиться так, что каждая новая реализация будет давать кластеризацию, которая будет существенно отличаться от предыдущей (по составу кластеров). В настоящей работе устойчивость получаемой кластеризации определялась

в ходе вычислительного эксперимента: кластеризация считалась устойчивой, если из ста реализаций метода динамических ядер не менее 60 совпадали.

Следует подчеркнуть, что результаты построения кластеризации методом динамических ядер почти всегда чувствительны к составу анализируемой базы данных: малое изменение в составе баз — например, добавление малого (по сравнению с мощностью базы) числа новых точек, либо удаление малого числа точек — может приводить к очень существенным изменениям в результирующей кластеризации. Эта чувствительность является ещё одним видом устойчивости метода; в рамках настоящей работы было установлено, что те базы, для которых строилась кластеризация, были устойчивыми в этом смысле. Метод динамических ядер порождает наибольшее число различимых классов. В рамках настоящей работы различимость классов не проверялась.

2.1.5 Классификация снизу вверх и сверху вниз

Для целей анализа генетических данных по связи структуры (частотных словарей) и филогении можно использовать два способа построения линейной кластеризации —«сверху вниз» и «снизу вверх». В первом случае весь массив данных необходимо разбить на минимальное количество классов, которое даёт более менее устойчивое разбиение на классы (здесь делили на два). В дальнейшем базу продолжают последовательно делить до тех пор, пока не получится структура типа дерева (см. рис. 1.1). Второй способ состоит в том, чтобы также последовательно делить исходное множество на 2, 3, ..., L классов до тех пор, пока всё более или менее устойчиво делится. Затем прослеживается судьба геномов из j-го класса ($1 \leqslant j \leqslant R$) при переходе от разбиения на R классов к разбиению на $R-1$ класс; здесь $L = \max\{R\}$. В

настоящей работе использовался второй метод.

2.1.6 Устойчивая и неустойчивая кластеризация

Введём понятие устойчивости: пусть делается M реализаций метода динамических ядер с разными случайными начальными разбиениями геномов по классам. Будем говорить, что финальная кластеризация будет устойчивой, если результирующее распределение геномов по классам всегда одно и то же при различных начальных разбиениях.

Однако следует отличать устойчивое распределение, получаемое при работе с методом динамических ядер, и устойчивость по отношению к индексированию базы данных. Большая часть геномов, составляющих эту базу, чаще всего принадлежала одному роду. К примеру, база содержала около 30 геномов видов рода *Pinus*. Существенное смещение базы, проявляющееся в наличии большого числа очень близких точек, будет заметно искажать результаты кластеризации, получаемые методом динамических ядер. Как следствие, выявить какую-либо закономерность здесь практически невозможно. Исходная база геномов хлоропластов индексировалась: из базы исключались геномы, принадлежащие одному роду (семейству) таким образом, чтобы общее число геномов одного рода не превышало определённого порога. Исключение геномов проводилось случайным образом, порог исключения составлял 5 геномов — это означает, что один род (либо семейство) может представлять не более 5 видов.

Определение устойчивости, сформулированное выше, зависит от двух параметров:

– частота δ попадания в один и тот же класс геномов данной группы. По-

нятно, что такой показатель не имеет естественного определения: его требуется определять каждый раз заново. Один из вариантов — сравнить его с частотой случайного попадания одной группы геномов в один и тот же класс. В рамках настоящей работы мы полагали $\min\{\delta\} \sim 0{,}5$;

– мощность получаемого класса. Действительно, пусть всего требуется разбить K точек на T классов. В практически значимых случаях $T \ll K$. Это означает, что число элементов (геномов, в нашем случае), попадающих в один класс должно иметь порядок K/T.

2.2 Метод упругих карт

В программе VidaExpert построить вложенное многообразие данных можно с помощью технологии упругих карт.

Упругую карту можно сравнить с куском упругой пластинки. Такая пластинка при разных деформациях стремиться восстановить свою первоначальную форму. Деформировать пластинку можно двумя способами: растягивая её «вдоль» и изгибая «поперёк». В первом случае она стремится сохранить свою длину, во втором — свою плоскую форму. Силы, образующиеся в данных ситуациях, можно назвать соответственно упругостью по отношению к расстоянию и упругостью по отношению к изгибу. Чтобы сетка обладала такими свойствами, необходимо добавить меру суммарного растяжения сетки и меру суммарного изгиба. Кроме этих двух мер, упругая сетка обладает мерой среднего квадрата расстояния до узла.

У метода упругих карт есть пара недостатков. С одной стороны, чем менее упруга карта (чем легче её согнуть и/или растянуть), тем точнее она описывает данные, но при этом воспроизводятся и все случайные шумы, ко-

торые обычно присущи реальным данным, ухудшается способность модели к обобщению информации. С другой, чем карта более эластична, тем более гладкую модель данных она собой представляет, но и тем хуже она описывает малые отклонения от предполагаемой закономерности.

После того, как карта построена и все точки данных перенесены на её поверхность, можно переходить к двумерной системе координат и работать уже с двумерной картой. Все исследования в настоящей работе проводились на плоской карте.

Глава 3. Результаты и обсуждение

3.1 Результаты кластеризации геномов хлоропластов

Результатом кластеризации методом динамических ядер стало выделение кластеров. При этом, чем меньше было число классов, на которое разбивают весь массив данных, тем более устойчивыми получались кластеры. В данном случае, устойчивыми считались те кластеры, в которых группа геномов на протяжении всего разбиения не меняла (или меняла очень редко) свой состав, всегда «вместе» попадая в один и тот же кластер. В настоящей работе делалось всего 100 реализаций. В программе Excel проводилось сравнение кластеров. Проверялись две устойчивости в распределении: одна по сумме реализаций — кластеры относились к устойчивым, если сумма их реализаций была равной 100, 99, 98 и 0, 1, 2. Последние (0, 1, 2) учитываются, потому что эти геномы лежат в переходе между кластерами.

Вторая устойчивость определялась эмпирическим путём. Все реализации просматривались и из них выделялись те, что совпадали. В итоге оказалось, что таких реализаций 60 из 100. К примеру, при делении на два класса, группа геномов чётко делилась на два класса и сколько бы не делалось дальнейших реализаций, состав классов оставался постоянным. При делении на восемь классов кластеры выделялись менее регулярно — часть всего массива данных постоянно перемещалась по кластерам. Геномы не меняющие класте-

ры на протяжении всего распределения так же выделялись.

Так же следует обратить внимание на то, что в некоторых случаях наблюдались геномы, которые в зависимости от количества классов входили то в группу «устойчивых» геномов, то в группу «неустойчивых». Иными словами, при делении на K классов какая-то группа геномов образует кластер и на протяжении всех 100 реализаций метода динамических ядер остаётся тем же составом, но при делении на $K + 1$ классов эти геномы по каким-то причинам начинать менять кластеры, присоединяясь то к одной, то к другой группе. Далее (не всегда, иногда группа геномов окончательно уходила в неустойчивую часть), при делении на $K + 2$ эта группа геномов опять занимает устойчивое положение.

В общем случае, устойчивость классификации методом динамических ядер отнюдь не гарантирована: мы можем столкнуться с ситуаций, в которой разбиение (например) на четыре класса оказывается весьма устойчивым в смысле попадания большого числа геномов в один и тот же класс, а разбиение на пять классов — неустойчивым. При этом разбиение на шесть (например) классов может опять оказаться устойчивым.

Один из подходов к решению указанной проблемы состоял в следующем: определять разделимость классов и объединять неразличимые классы.

3.1.1 Различимость классов

Введём понятие различимости классов. Классы считались различимыми, если их радиусы не больше расстояния между их цетрами. В связи с этим также различают два вида различимости: хорошую и плохую. Классы хорошо различимы, если сумма их радиусов не больше расстояния между их

центрами, и плохо различимы, если расстояние между центрами классов не меньше одного из больших радиусов.

В данной работе исследования разделимости (она же различимость) классов не проводилась. Однако, исследовалось включение классов при переходе от заданной классификации и той, у которой задано меньшее число классов. То есть, начиная с какого-нибудь «меньшего» числа классов, добивались для него устойчивой кластеризации. Затем увеличивали число классов и следили за тем, в какие из классов, представленных в новой классификации с большим числом классов, попадают те геномы, которые ранее составляли «ушедший» класс.

3.2 О составе классов, выделяемых методом динамических ядер

При кластеризации методом динамических ядер весь массив данных разбился на две части: часть геномов, которые делились устойчиво и часть, которые делились неустойчиво. Устойчивость здесь понимается как то обстоятельство, что состав порождаемых классов всегда был одним и тем же, вне зависимости от начального[1] разбиения геномов по классам. Соответственно, отсутствие устойчивости, наблюдавшееся для части геномов, означает, что их принадлежность к тому или иному классу менялась в разных реализациях метода.

Разберём оба случая более подробно. На рис. 3.1 и 3.2 представлена та часть геномов, которые распределялись устойчиво. Как уже было сказано выше, в настоящей работе использовался способ построения классификации

[1]Напомним, что начальное разбиение точек по классам всегда является случайным.

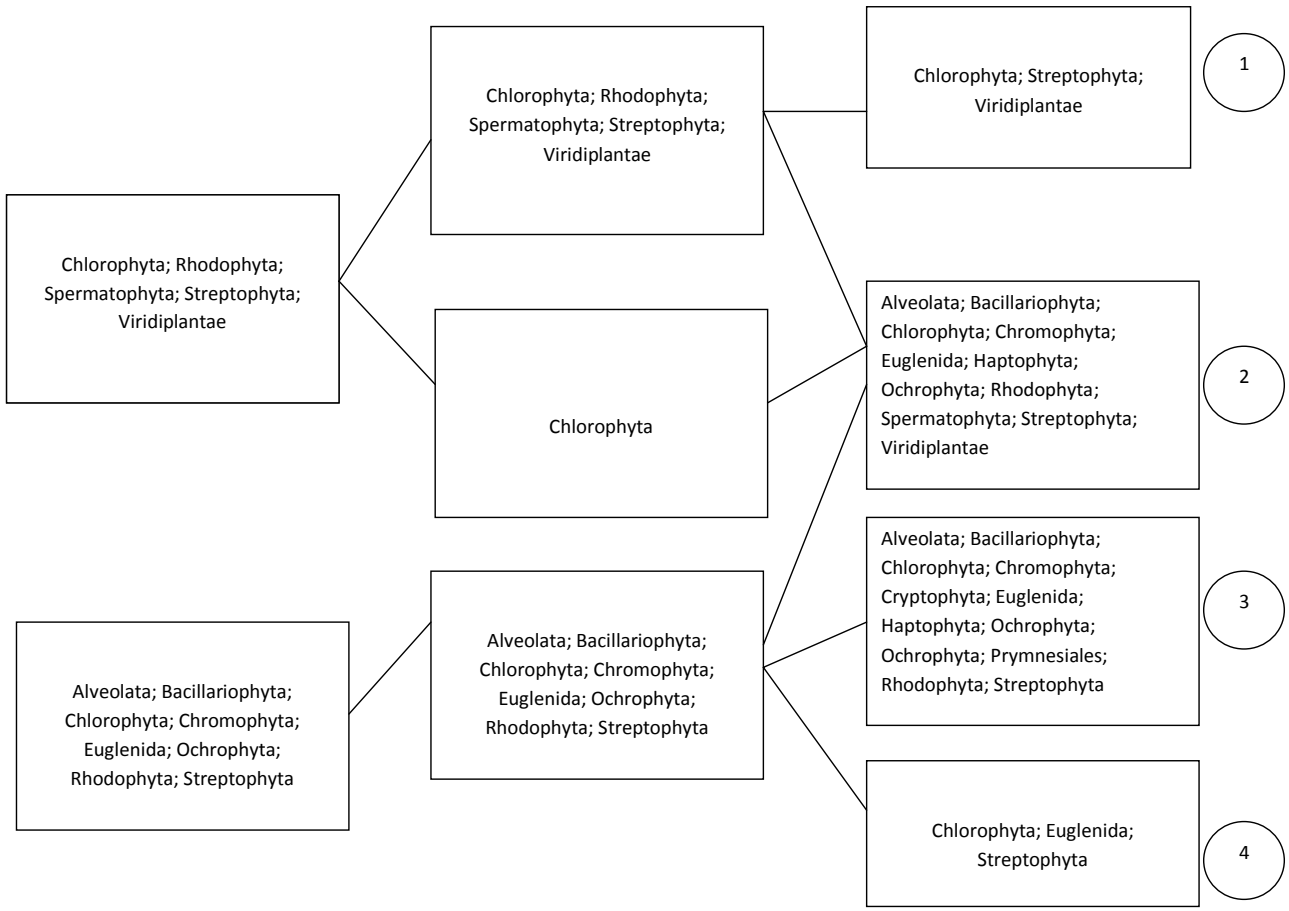

Рис. 3.1. Устойчивое распределение геномов высших растений, слои с 1-го по 3-ий.

«снизу вверх». Сначала общую базу разбили на два класса. Далее всю базу делили последовательно на три, четыре и так до восьми классов. При переходе от одного слоя классов к другому прослеживалась судьба каждого генома и каждого класса. Важно было определить, каким образом меняются составы классов для того, чтобы выявить закономерность переходов геномов от одного классу к другому, если она есть. Геномы вели себя весьма структурированно как при устойчивом распределении, так и при неустойчивом. В первом случае чётко выделялись группы геномов, которые составляли один кластер или два (три) близ лежащих на протяжении всей кластеризации. В свою очередь группа геномов, которую приписали неустойчивой классифи-

кации, напротив, меняла свои классы, переходя то к одному кластеру, то к другому. При этом, каждая такая группа имела определённый состав, который не менялся. Иными словами, кластеры меняли не единичные геномы, а группы.

Ниже представлены полученные результаты: устойчивое распределение и неустойчивое. Следует добавить, что каждое распределение представлено двумя картинками. На самом деле обе картинки представляют одно распределение, но разделены на две для удобства восприятия. Цифры в кружочках в конце первого рисунка и начале второго являются в некотором смысле рёбрами между слоями (показывают переход от класса большего уровня к классу меньшего уровня).

В настоящей работе проверялась связь между структурой и таксономией. Физически они друг с другом никак не связаны. Близость в смысле структуры определялась по частотным словарям. В свою очередь филогенетическая близость определялась по соматическому геному.

Важно, что кластеризация проводилась по структуре, а анализ результатов по таксономии.

При делении на два класса было замечено, что все геномы разделились на высшие растения: сюда попали голосеменные и покрытосеменные и низшие — здесь большую часть занимают водоросли. Тем не менее, весьма интересно повели себя растения подцарства *Chlorophyta*, которые разделились на две группы. В первую попали *Chlorella*, *Chlamydomonas*, *Ctenocladaceae*, *Pycnococcus*, *Pedinomonas*, *Aureoumbra*, *Auxenochlorell*, *Ostreococcus*, *Chlorella*, *Prasinococcus*, *Prasinoderma*, *Prasinophytes*, *Dictyochloropsis*, *Myrmecia*, *Stichococcus*, *Chlorella*, *Pabia*, *Pedinomonas*, *Pedinomonas*, *Neocystis*, *Micro-*

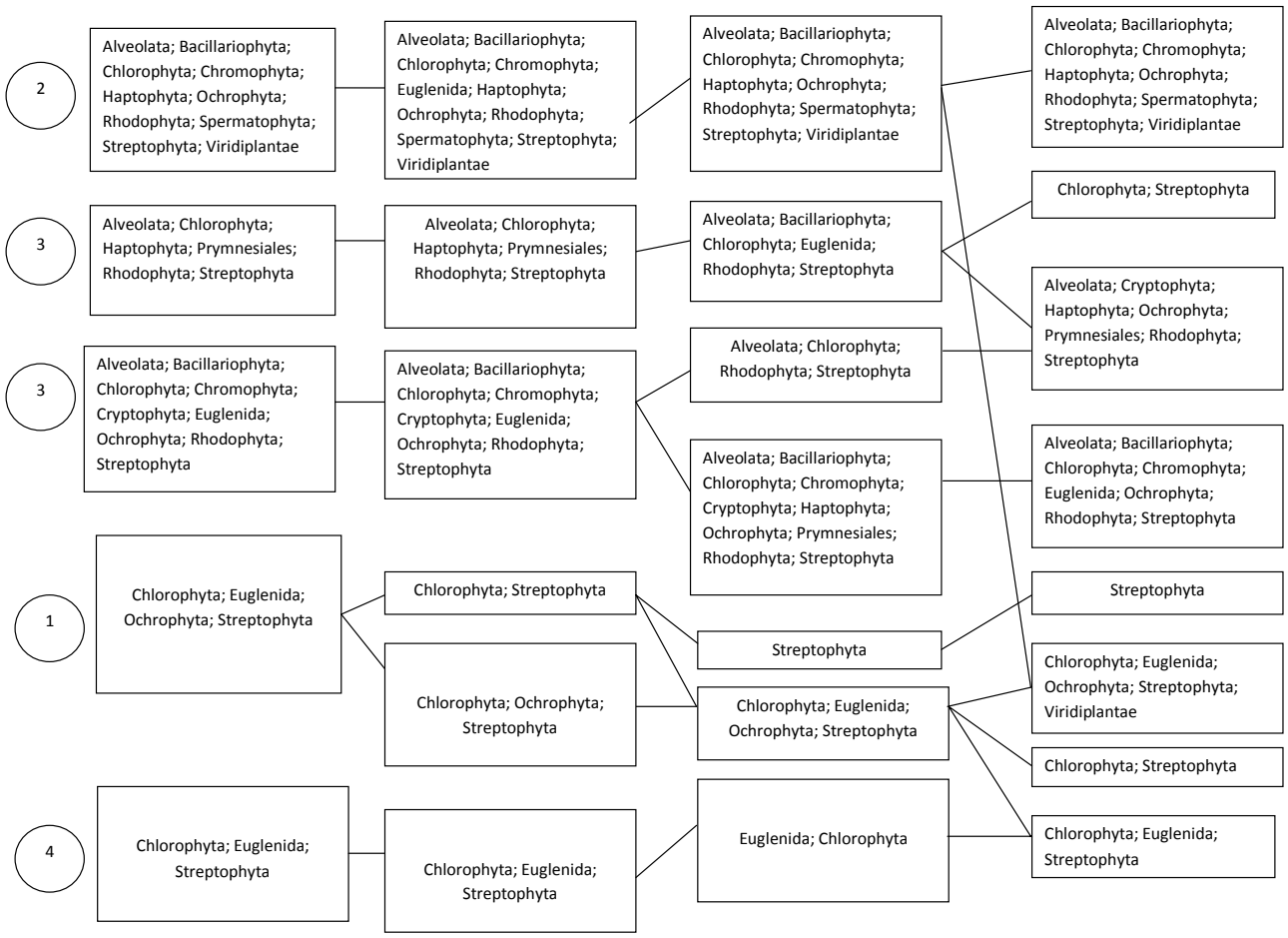

Рис. 3.2. Устойчивое распределение геномов высших растений, слои с 4-го по 7-ой.

thamnion, *Lauterbornii*, *Fusochloris*, *Dicloster*, во второй оказались *Festuca*, *Typha*, *Berberis*, *Parasitica*. Кроме того, при делении на три класса выделилась группа геномов организмов, также принадлежащих подцарству *Chlorophyta*; в её состав входили *Micromonas*, *Selaginella*, *Ostreococcus*, *Prasinophytes*, *Parasitica* и *Multiseta*.

Не меньший интерес вызвала к себе группа геномов, что делилась неустойчиво. Группы геномов меняли классы структурированно. Прежде всего они также объединились в группы и меняли классы именно этим составом. На рис. 3.3 и 3.4 можно увидеть состав этих классов. Причины, объясняющие такое поведение этих геномов может быть их близкородственные связи.

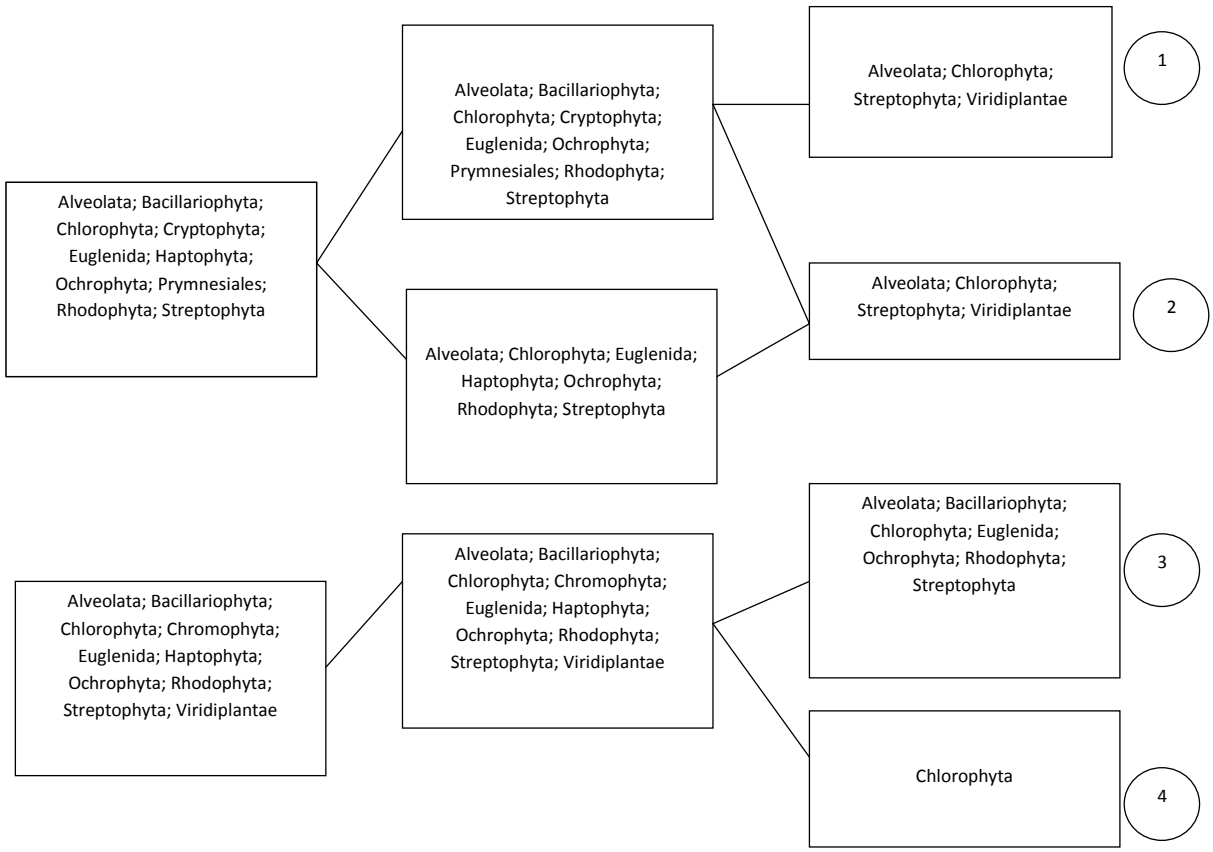

Рис. 3.3. Неустойчивое распределение геномов высших растений

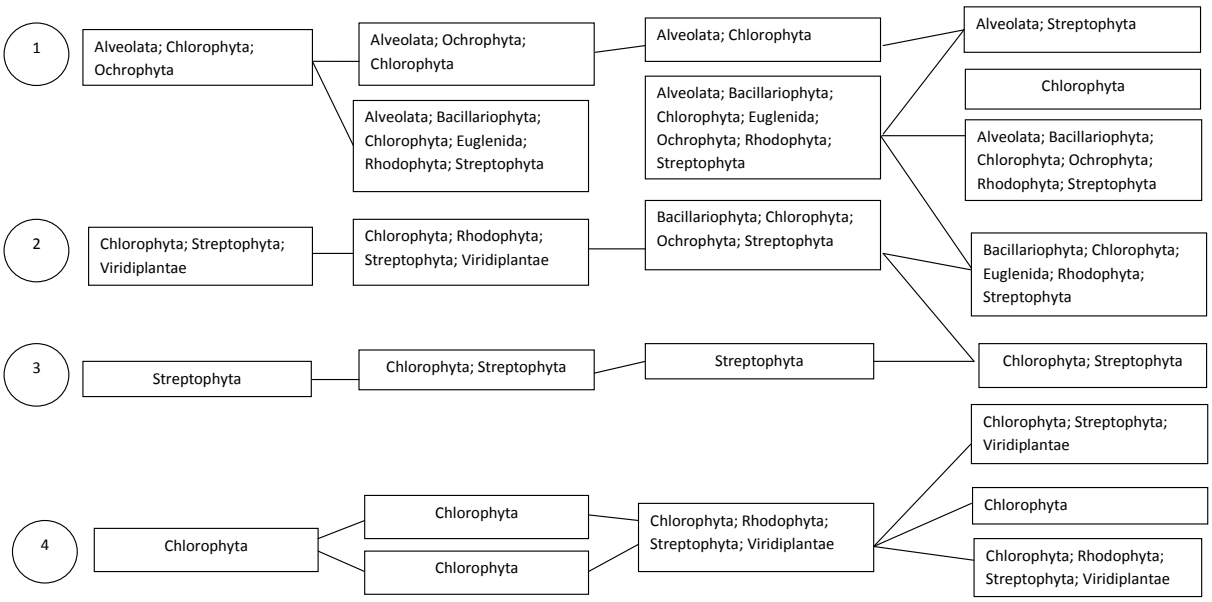

Рис. 3.4. Неустойчивое распределение геномов высших растений

Глава 4. Выводы

Высоко упорядоченное распределение видов и родов по классам, определяемым лишь частотами триплетов в геномах хлоропластов, доказывает факт сильной синхронизации в эволюции двух генетических систем — соматической и геномов хлоропластов.

Таким образом, показано существование очень высокого уровня синхронизации геномов хлоропластов и соматических геномов растений, несущих эти хлоропласты. Физически они друг с другом никак не связаны. Доказательством служит тот факт, что различные таксоны расходятся по классам неслучайно: выделяются весьма устойчивые группы родов/семейств, всегда попадающие в один класс.

Положения, выносимые на защиту

1. Предложен метод, позволяющий строить филогенетические деревья для различных видов по всему геному в целом, не выделяя в нём отдельных специфических областей (например, группы консервативных генов).
2. Показано существование высокого уровня синхронизации геномов хлоропластов и соматических геномов растений, несущих эти хлоропласты. Физически они связаны друг с другом слишком слабо, чтобы можно было ожидать наблюдаемой синхронии как результата такого взаимодействия.
3. Наблюдаемая кластеризация обладала высокой устойчивостью. При этом группа геномов, проявлявшая неустойчивость при кластеризации, также обладала специфической устойчивостью: они меняли свою принадлежность к тому или иному кластеру в ходе построения классификации, однако такие переходы осуществлялись регулярно. Иными словами, группа геномов, проявлявших неустойчивость, перераспределялась из класс в класс так же согласованно, не делясь на более мелкие подгруппы.

Литература

[1] Foulongne-Oriol M., Murat C., Castanera R., Ramirez L., Sonnenberg A. S. Genome-wide survey of repetitive DNA elements in the button mushroom Agaricus bisporus. Fungal Genet Biol. 2013; 55: 6-21.

[2] Sharma M. K., Sharma R., Peijian Cao, Jenkins J., Bartley L. E., Qualls M., Grimwood J., Schmutz J., Rokhsar D., Ronald P. C. A Genome-Wide Survey of Switchgrass Genome Structure and Organization. 2012.

[3] Fischer N. O., Tok J. B., Tarasow T. M. Massively parallel interrogatooin of aptamer sequence, topology and structure. Nucleic Acids Res. 2006

[4] http://fizrast.ru/fotosintez/hloroplasty/stroenie.html

[5] Tiwari A.K., Srivastava R.: A survey of computational intrlligence techniques in protein function prediction. Int. J. Proteomics. 2014.

[6] Provata A., Nicolis C., Nicolis G., DNA viewed as an out-of-equilibrium structure. Phys.Rev. E 89, 2014.

[7] Остин Оре. 1980. Теория графов.22–30, 34–51.

[8] Р. В. Гнутова «Современные тенденции в таксономии и номенклатуре вирусов»

[9] Б. Б. Бадмаев «Сравнение таксономической структуры кормовых растений двух видов наземных беличьих западного забайкалья и степной флоры региона»

[10] A. Batko «Phylogenesis and taxonomic structure of the entomophthoraceae»

[11] W. K. Taia «Modern trends in plant taxonomy»

[12] Зиновьев А.Ю. Визуализация многомерных данных, 65–74.

[13] Горбань А.Н., Россиев Д.А. 1996. Нейронные сети на персональном компьютере, 114–232.

[14] M.G. Sadovsky, N.A. Zaitseva, Yu.A. Putintseva. 2011. System biology on mitochondrion genomes. Biotechno. 61-66

[15] Gorban A.N., Popova T.G., Sadovsky M.G., Wunsch D.C. 2001. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts. *Intelligent Engineering Systems through Artificial Neural Networks* **11** — *Smart Engineering System Design*, N.-Y.: ASME Press, 2001, 657–663.

[16] Горбань А.Н., Попова Т.Г., Садовский М.Г. 2003. Классификация нуклеотидных последовательностей по частотным словарям обнаруживает связь между их структурой и таксономическим положением организмов. *Журн.общей биол.* **64**, 51–63.

[17] Sadovsky M.G., Shchepanovsky A.S., Putintzeva Yu.A. Genes, Information and Sense: Complexity and Knowledge Retrieval // *Theory in Biosciences*, 2008, **127**, pp. 69 – 78.

[18] М. А. Мамонтова и М. Г. Садовского «Информационная ценность различных триплетовнекоторых генетических систем»

[19] Fukunaga K., 1990/ Introduction to statistical patten recognition. Academic Press: London. 591 p.

[20] Sadovsky M.G., 2002. On the redundancy of viral and prokaryotic genomes. **5**, 695-701.

[21] http://intellect-tver.ru/?p=265

[22] https://habrahabr.ru/post/67078/

[23] Yu. Putinzeva, A. Chernyshova, V. Fedotova. 2015. IWBBIO Genome Structure of organelles strongly relates to taxonomy of bearers // LNCS, vol.9044, Part II, pp.481 – 490.

[24] Чернышова А.И. 2015, Синхронизация эволюции растений и их хлоропластов // Труды XIV Межд. ФАМ-конференции, Красноярск, изд-во СФУ.

[25] Садовский М.Г., Чернышова А.И. 2014. Проявление синхронизации в эволюции геномов растений и их хлоропластов. 148–152.

[26] Садовский М.Г., Чернышова А.И. 2015. ВЫЯВЛЕНИЕ СВЯЗИ МЕЖДУ СТРУКТУРОЙ И ТАКСОНОМИЕЙ ГЕНОМОВ ХЛОРОПЛАСТОВ // Международная научная конференция «Проспект Свободный — 2015», Красноярск.

[27] Садовский М.Г., Чернышова А.И. 2016. СВЯЗЬ МЕЖДУ СТРУКТУРОЙ И ТАКСОНОМИЕЙ ГЕНОМОВ ХЛОРОПЛАСТОВ ХВОЙНЫХ // IX сибирский конгресс женщин-математиков, Красноярск.

[28] Садовский М.Г., Чернышова А.И. 2014. ВЫЯВЛЕНИЕ СВЯЗИ СТРУКТУРЫ И ТАКСОНОМИИ ГЕНОМОВ ХЛОРОПЛАСТОВ МЕТОДОМ ДИНАМИЧЕСКИХ ЯДЕР // Фундаментальные исследования, Москва, С. 545-549.

[29] Садовский М.Г., Чернышова А.И. 2016. Построение связи между структурой и таксономией геномов хлоропластов сосен // Международная научная конференция «Проспект Свободный — 2016», Красноярск.

[30] МНСК 2015, «Проявление синхронизации в эволюции геномов растений», г. Новосибирск;

[31] Всероссийский семинар по нейроинформатике, «Проявление синхронизации в эволюции геномов растений и их хлоропластов», 2014 г., г. Красноярск;

[32] МНСК 2014, «Выявление связи между структурой и таксономией геномов хлоропластов», г. Новосибирск.

Список таблиц

Список иллюстраций